

METHODS AND APPARATUS FOR ESTIMATING SIMILARITY

BACKGROUND OF THE INVENTIONA. Field of the Invention

[0001] The present invention relates generally to similarity estimation, and more particularly, to calculating similarity metrics for objects such as web pages.

B. Description of Related Art

[0002] The World Wide Web ("web") contains a vast amount of information. Locating a desired portion of the information, however, can be challenging. Search engines catalog web pages to assist web users in locating the information they desire. Typically, in response to a user's request, the search engine returns references to documents relevant to the request.

[0003] From the search engine's perspective, one problem in cataloging the large number of available web pages is that multiple ones of the web documents are often identical or nearly identical. Separately cataloging similar documents is inefficient and can be frustrating for the user if, in response to a request, a list of nearly identical documents is returned. Accordingly, it is desirable for the search engine to identify documents that are similar or "roughly the same" so that this type of redundancy in search results can be avoided.

[0004] In addition to improving web search results, the identification of similar documents can be beneficial in other areas. For example, storage space may be reduced by storing only one version of a set of similar documents. Or, a

collection of documents can be grouped together based on document similarities, thereby improving efficiency when compressing the collection of documents.

[0005] One conventional technique for determining similarity is based on the concept of sets. A document, for example, may be represented as a sub-set of words from a corpus of possible words. The similarity, or resemblance of two documents to one another is then defined as the intersection of the two sets divided by the union of the two sets. One problem with this set-based similarity measure is that there is limited flexibility in weighting the importance of the elements within a set. A word is either in a set or it is not in a set. In practice, however, it may be desirable to weight certain words, such as words that occur relatively infrequently in the corpus, more heavily when determining the similarity of documents.

[0006] Accordingly, there is a need in the art for improved techniques for determining similarity between documents.

SUMMARY OF THE INVENTION

[0007] Systems and methods consistent with the present invention address this and other needs by providing a similarity engine that generates compact representations of objects that can be compared to determine similarity between the objects.

[0008] In one aspect, the present invention is a method for generating a compact representation of an object. The method includes generating a vector corresponding to the object, each coordinate of the vector being associated with a

corresponding weight and multiplying the weight associated with each coordinate in the vector by a corresponding hashing vector to generate a product vector. The method further includes summing the product vectors and generating the compact representation of the object using the summed product vectors.

[0009] A second method consistent with the present invention includes creating a similarity sketch for each of first and second objects based on the application of a hashing function to a vector representation of the first and second objects. Additionally, the method compares, on a bit-by-bit basis, the similarity sketches for the first and second objects, and generates a value defining the similarity between the first and second objects based on a correspondence in the bit-by-bit comparison.

[0010] Another aspect of the present invention is directed to a server that includes at least one processor, a database containing a group of objects, and a memory operatively coupled to the processor. The memory stores program instructions that when executed by the processor, cause the processor to remove similar objects from the database by comparing similarity sketches of pairs of objects in the database and removing one of the objects of the pair when the comparison indicates that the pair of objects are more similar than a threshold level of similarity. The processor generates the similarity sketches for each of the pair of objects based on an application of a hashing function to vector representations of the objects.

[0011] Yet another aspect of the invention is directed to a method for generating a compact representation of a first object. The method includes

identifying a set of features corresponding to the object and generating, for each feature, a hashing vector having n coordinates. The hashing vectors are summed to obtain a summed vector, and an $n \times x$ -bit representation of the summed vector is obtained by calculating an x -bit value for each coordinate of the summed vector.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate an embodiment of the invention and, together with the description, explain the invention. In the drawings,

[0013] Fig. 1 is a diagram illustrating an exemplary system in which concepts consistent with the present invention may be implemented;

[0014] Fig. 2 is a flow chart illustrating methods consistent with principles of the present invention for generating similarity sketches;

[0015] Figs. 3A and 3B illustrate exemplary object vectors; and

[0016] Fig. 4 is a diagram conceptually illustrating the calculation of a result vector for an exemplary object vector based on two corresponding hashing vectors.

DETAILED DESCRIPTION

[0017] The following detailed description of the invention refers to the accompanying drawings. The detailed description does not limit the invention. Instead, the scope of the invention is defined by the appended claims and equivalents.

[0018] As described herein, a similarity engine generates compact representations, called similarity sketches, for object vectors. The object vectors can be created for objects, such as document files. The similarity sketch for two different object vectors can be quickly and easily compared to generate an indication of the similarity between the two objects.

[0019] Fig. 1 is a diagram illustrating an exemplary system in which concepts consistent with the present invention may be implemented. The system includes multiple client devices 102, a server device 110, and a network 101, which may be, for example, the Internet. Client devices 102 each include a computer-readable medium 109, such as random access memory, coupled to a processor 108. Processor 108 executes program instructions stored in memory 109. Client devices 102 may also include a number of additional external or internal devices, such as, without limitation, a mouse, a CD-ROM, a keyboard, and a display.

[0020] Through client devices 102, users 105 can communicate over network 101 with each other and with other systems and devices coupled to network 101, such as server device 110.

[0021] Similar to client devices 102, server device 110 may include a processor 111 coupled to a computer readable memory 112. Server device 110 may additionally include a secondary storage element, such as database 130.

[0022] Client processors 108 and server processor 111 can be any of a number of well known computer processors, such as processors from Intel Corporation, of Santa Clara, California. In general, client device 102 may be any

type of computing platform connected to a network and that interacts with application programs, such as a digital assistant or a "smart" cellular telephone or pager. Server 110, although depicted as a single computer system, may be implemented as a network of computer processors.

[0023] Memory 112 may contain a number of programs, such as search engine program 120, spider program 122, and similarity engine 124. Search engine program 120 locates relevant information in response to search queries from users 105. In particular, users 105 send search queries to server device 110, which responds by returning a list of relevant information to the user 105. Typically, users 105 ask server device 110 to locate web pages relating to a particular topic and stored at other devices or systems connected to network 101. The general implementation of search engines is well known in the art and therefore will not be described further herein.

[0024] Search engine program 120 may access database 130 to obtain results from a document corpus 135, stored in database 130, by comparing the terms in the user's search query to the documents in the corpus. The information in document corpus 135 may be gathered by spider program 122, which "crawls" web documents on network 101 based on their hyperlinks.

[0025] In one embodiment consistent with the principles of the present invention, memory 112 additionally includes similarity engine 124. Similarity engine 124 generates similarity information between documents in document corpus 135. More particularly, similarity engine 124 may generate a relatively small "sketch" (e.g., a 64 bit value) for each document in corpus 135 and may

compare the sketches between pairs of documents. If two documents, based on their sketches, are more similar than a predefined similarity threshold, similarity generator 124 may, for example, remove one of the two documents from document corpus 135. In this manner, search engine program 120 is less likely to return redundant or overly similar results in response to user queries. Alternately, instead of removing documents from the corpus based on the similarity sketches, similarity generator 124 may remove duplicate documents "on the fly." In other words, search engine program 120 may generate a set of documents in response to a user query and similarity generator 124 may cull the generated set of documents based on similarity sketches.

[0026] Fig. 2 is a flow chart illustrating methods consistent with principles of the present invention for generating similarity sketches for objects by similarity engine 124. In general, similarity engine 124 processes objects, such as web documents, to generate the sketches. Sketches of different objects can be compared to generate a quantitative metric of the similarity of the objects.

[0027] To begin, similarity engine 124 creates a vector representation of the input object. (Act 201). The input object may be a document such as a web page. More generally, the object can be any item that contains a series of discrete elements, and the representation of the input object can be a set of associated features rather than a vector per se. The vector coordinates are indexed by the elements of the object. In the case of documents, for example, the elements may be the words in the document.

[0028] Fig. 3A illustrates an example of a vector 301 for the object phrase "four score and seven years ago." The vector for this phrase contains six non-zero coordinates corresponding to the six words in the phrase. Conceptually, the vector can be thought of as containing as many coordinates as there are elements in the universe of elements (e.g., the number of unique words in the corpus), with each coordinate other than those in the phrase having a value of zero.

[0029] In Fig. 3A, each non-zero coordinate is given an equal vector weight (i.e., it has the value one). More generally, however, different elements can be given different weights. This is illustrated in Fig. 3B, in which words that are considered "more important" are more heavily weighted in object vector 302. For example, "years" and "score," because they are less common words, may be given a higher weighting value, while "and" is given a low weight. In this manner, the object vector can emphasize certain elements in the similarity calculation while de-emphasizing others. In the implementation of Fig. 3B, valid values are in the range between zero and one.

[0030] The weights for an element can be calculated using any number of possible weighting formulas. In one implementation, for vectors associated with documents, the weight for a word is set as the number of occurrences of the word within the document divided by the number of documents in the corpus containing the word.

[0031] After constructing the object vector 301 or 302 for the object, similarity engine 124 multiplies every coordinate in the object vector by a

corresponding predetermined hashing vector having a preset number of coordinates (dimensions), where each coordinate in the hashing vector corresponds to a number, such as a random number, provided that the same input object vector coordinate yields the same hashing vector coordinate. (Act 202). A different hashing vector may be used for each possible coordinate (e.g., word). The number of preset coordinates in the hashing vector corresponds to the size of the resulting sketch. For example, to produce a 64-bit sketch for each object, a 64-dimensional hashing vector would be chosen.

[0032] For a particular input vector coordinate, the corresponding hashing vector could be generated from a pseudo random number generator seeded based on the input vector coordinate (ignoring the actual value of the input vector coordinate). The values of the coordinates of the hashing vector could be selected based on the output of the pseudo random number generator so that the values are random numbers drawn from a chosen distribution. For example, coordinates could be drawn from the normal distribution with probability density function $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Another possibility is to choose coordinates to be either +1 or -1 with equal probability. A hash function could be used to map input vector coordinates to hashing vectors directly. For example, a 64-bit hash value obtained from a hash function could be mapped to a 64-dimensional hashing vector by choosing the i^{th} coordinate of the hashing vector to be +1 or -1 based on whether the i^{th} bit in the hash value is 1 or 0.

[0033] The product vectors generated in Act 202 are summed for all elements in the object to generate a result vector having the same number of coordinates as each of the predetermined hashing vectors. (Act 203).

[0034] Stated more formally, the operations performed in Acts 202-203 generate the result vector based on a hashing function defined as,

$$\bar{R} = \sum_i c_i \bar{v}_i$$

where \bar{R} is the result vector, c_i is the weight value for the i^{th} coordinate (a scalar), \bar{v}_i is the predetermined hashing vector for the i^{th} coordinate in the object vector, and the sum is taken over all the possible coordinates in the object vector.

[0035] The result vector can be used as the similarity sketch for the object. Optionally, the result vector may be further simplified to obtain the similarity sketch by generating, for each coordinate in the resultant vector, a single bit corresponding to the sign of the coordinate's value. (Act 204). The generated bits are then concatenated together to form the object's sketch. (Act 205). In the above example, for a 64-coordinate hashing vector, and thus, a 64-coordinate result vector, the similarity sketch would be a 64-bit value.

[0036] When comparing the similarity sketches for two objects generated by Act 205, similarity generator 124 may create a similarity value based on a bit-by-bit comparison of the sketches. In one implementation, a final similarity measure between two sketches can be obtained by logically exclusive-ORing (XOR) the two sketch values and summing the complements of the individual result bits to yield a value between zero (not similar) and 64 (very similar) for a 64-bit similarity sketch. If the similarity measure is above a predetermined

threshold (e.g., greater than 50), the two objects may be considered to be similar to one another.

[0037] Fig. 4 is a diagram conceptually illustrating the calculation of a result vector for an exemplary object vector 401 and two corresponding hashing vectors 402 and 403. In practice, object vectors are likely to be significantly more complex than object vector 401 and the hashing vectors are likely to contain more than the three coordinates illustrated in hashing vectors 402 and 403.

[0038] Object vector 401 contains two non-zero elements, "four" and "score," each associated with a weight value. Hashing vectors 402 and 403 each have three coordinates, which will thus yield a three-bit similarity sketch. Result vector 404 is the sum of hashing vectors 402 and 403 after being multiplied by the value of the corresponding coordinate in object vector 401. Similarity sketch 405 is derived from the signs of the three coordinates in result vector 404. As shown, because each of the three coordinate values is positive, the similarity sketch is "111."

[0039] The values of the coordinates in hashing vectors 402 and 403 can be chosen at random. For a large corpus (e.g., a million possible words) and a larger similarity sketch (e.g., 128 bits), such a large number of hashing vectors (i.e., one million), each made up of 128 random values, can be burdensome to store and subsequently access. Accordingly, the random numbers for each hash vector could be dynamically generated using a pseudo random number generator seeded with a value based on the hash vector's coordinate. Furthermore, if it is determined that two coordinates in the object vector are "similar", it may be

desirable to make the hashing vectors similar for these two object vector coordinates. In this manner, one may incorporate information about the similarity of object vector coordinates into their associated hashing vectors.

[0040] Although the above description of similarity engine 124 generally describes a similarity engine operating in the context of web documents, the concepts described could also be implemented based on any object that contains a series of discrete elements.

[0041] As mentioned, similarity sketches generated by similarity engine 124 may be used to refine the entries in database 130 to reduce the occurrence of redundant or nearly redundant documents returned in response to a user's search query. The similarity sketches can be used in a number of other applications. For example, spider program 122, when crawling web sites, may use similarity engine 124 to determine sites that are substantial duplicates of one another. Mirror sites are one example of duplicate sites that occur frequently on the web. In future crawls, spider program 122 may more efficiently crawl the web by avoiding crawling the sites that are determined to be substantial duplicates.

[0042] As another example of the application of similarity engine 124, similarity engine 124 may generate object vectors for a web document based on the list of hyperlinks in the document. Accordingly, similarity engine 124 may develop similarity sketches based on a document's list of links.

[0043] As yet another example of the application of similarity engine 124, the similarity engine may operate on "snippets" of text associated with documents returned from a search query. A snippet is an abstraction of the document that is

initially returned to a user in response to a search query. Based on the snippets, users select full documents to view. Similarity engine 124 may compare snippets from a search query and exclude those snippets and/or search results that exceed a similarity threshold.

[0044] As yet another example of the application of similarity engine 124, similarity engine 124 may quantitatively determine the similarity between two words or phrases based on the context surrounding those words or phrases from the document corpus. More specifically, similarity engine 124 may create an object vector for a word, such as "shop", based on the words within a certain distance (e.g., 10 words) of each occurrence of shop in the corpus. The coordinates of this object vector would be all the words in the corpus. The value for the coordinates may be, for example, the number of times the coordinate word occurs within the context distance. If, for example, the word "cart" occurs frequently in the document corpus within the selected context distance to the word "shop," then the value for "cart" would be relatively high in the object vector. Two different words with close similarity sketches, as determined by similarity engine 124, are likely to be used as synonyms in the corpus.

[0045] Similarity engine 124 may also be used outside of the context of web search engines. For example, a collection of objects may be grouped based on their similarity. Grouping objects in this manner may be used to gain increased compression ratios for the objects or to increase access speed from a storage medium.

[0046] The similarity sketches described above have a number of advantages. Because the similarity sketches are based on vector representations, different elements of the vectors can be given different weights, thus allowing vector elements to be more or less important than one another. In contrast, in conventional set-based similarity techniques, elements are simply either in a set or not in the set. Additionally, the sketches generated above are relatively compact compared to similarity measures generated with the conventional set-based techniques.

[0047] The foregoing description of preferred embodiments of the present invention provides illustration and description, but is not intended to be exhaustive or to limit the invention to the precise form disclosed. Modifications and variations are possible in light of the above teachings or may be acquired from practice of the invention. For example, although the preceding description generally discussed the operation of search engine 120 in the context of a search of documents on the world wide web, search engine 120 and similarity engine 124 could be implemented on any corpus. Moreover, while a series of acts have been presented with respect to Fig. 2, the order of the acts may be different in other implementations consistent with the present invention.

[0048] No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly described as such. Also, as used herein, the article "a" is intended to include one or more items. Where only one item is intended, the term "one" or similar language is used.

[0049] The scope of the invention is defined by the claims and their equivalents.